Determining Biases within Wikipedia Articles

Bret Duntley
Yaozong Huang
Saika Islam
Amelia Rave
Jordan Savage

1 Project Description

Wikipedia.org is a well-known free and public "modern encyclopedia" from which millions of people digest, learn, and share information. In contrast to classically-written encyclopedias that are authored and edited over many years by selected and qualified intellectuals, Wikipedia articles can be authored and edited by almost anyone. In this difference lies the potential for contamination of information by the personal biases and opinions held by such an author. With this in mind, the primary goal of our project is to determine whether Wikipedia as a platform portrays a given notable person as "famous" or "infamous" based on the language used to describe his or her historical actions. Wikipedia is a site that contains articles on nearly every historically significant topic and person (both living and deceased) and can maintain such large volumes of information by allowing subsets of users to collaboratively construct articles attempting to encompass all aspects of the person or topic from a neutral perspective. However, as is true with all human-constructed and subjective information regarding recollection and portrayal of historical events, there is an inherent bias in these articles. Wikipedia has even been publicly criticized for the production of biased articles. Specifically, Wikipedia was criticized for having an overall progressive slant to its information, due to a large number of its users being Democrats (Ashtear and Tezuka, 2020). We expect such a bias to be amplified when an article's content has the potential to intersect with political or otherwise controversial information. Given that a Wikipedia article is likely to have been constructed by multiple authors who come from different ideological backgrounds, one might expect there to be a convergence towards producing text documents that encompass the general populations' consensus around a person or topic. Using Wikipedia.org articles as both our training and test data, our program aims to determine whether a given notable person of the past or present is accurately presented as either famous or infamous according to a ground truth of Ranker lists.

2 Related Work

Researching similar works, it is evident that many others have attempted to analyze this phenomenon and the degree to which it takes place. Currently, there exists substantial research on the political, cultural, and gender bias put forth by different violations of the expected neutral point of view on which we intend to build. As was mentioned previously, Wikipedia was criticized for having an overall progressive slant to its information, due to a large number of its users being Democrats (Ashtear and Tezuka, 2020). Along with the Democratic slant itself, some of these works have also analyzed how Wikipedia articles' points of view have changed over time due to the increasing traffic on the web (Hube, 2017). According to researchers Sanmay Das and Adam Lavoie at Washington Univer-

sity in St. Louis, the number of editors for a given Wikipedia article has been decreasing since the creation of the shared database (Das and Lavoie, 2014). This prompts the question: do fewer editors point to more reliable and unbiased articles? The study found that this is true to a certain extent, but that articles on controversial topics are still predominantly written from a single point of view. Our project will investigate further into this 'extent' as it analyzes and categorizes articles that are currently on the decentralized platform, and thus have been edited or updated in recent years.

The study conducted by Das and Lavoi focuses on modeling topics and points of view in conjunction with one another using a Bayesian graph-based approach, unlike our model that fixes topics before considering points of view. By using this model, the researchers gained important insights on how specific articles were born and changed throughout their lifespan which, in turn, revealed information about the general functioning of the site and others like it. While our project focuses on these political, cultural, and gender biases and not point of view, we aim to build on these related works by considering the context of the creation of these articles and what may have contributed to the way it is read.

In the study, "Is Wikipedia Biased?" conducted by Shane Greenstein and Feng Zhu at Northwestern University, the authors investigated whether Wikipedia articles hold up to the model of "NPOV" or Neutral Point of View. They were able to measure the bias of about 40% of all Wikipedia articles and found that particularly controversial topics often did not have the bias as expected if the topic was popular enough(Greenstein and Zhu, 2012). Political topics like civil rights and trade were found to have Democratic and Republican biases, respectively. This considerable variance among the slant of a variety of topics made it difficult to generalize about bias and slant among all articles, but in general, the researchers found that older articles tended to lean Democratic, with more recent articles having less of a slant. Given this, it is easy to see how the number of revisions and the original date of publication would matter greatly for the articles we choose to investigate.

In a follow-up to this study, Greenstein and Zhu published a new work several years later at Harvard Business School titled: "Do Experts or Collective Intelligence Write with More Bias? Evidence from Encyclopædia Britannica and Wikipedia". This study focused on comparing the difference in bias between an expert-written(Encyclopædia Britannica) and a collectively written(Wikipedia) article. Surprisingly, these authors found that there was not a significant difference between the bias of these two sources, however, this was dependent on the Wikipedia article receiving a large number of edits and revisions(Greenstein and Zhu, 2016). It is important to note here that this conflicts with the findings of the Das and Lavoie study. Overall there was still a difference, as expected, in bias between Wikipedia articles and Encyclopædia Britannica articles, with Wikipedia being more biased. The authors surmise that this is largely due to the enormous amount of content Wikipedia has that can't undergo nearly the same level of editorial review and editing as Encyclopædia Britannica. This study highlights the reasons for our analysis of Wikipedia and helps support our desire to investigate sources of bias within Wikipedia articles, especially on controversial topics.

An assumption can be made that due to the vitriol and intensity of political discourse, articles about politics or political figures, some of whom we will likely investigate within our report, are more likely to contain bias. A study in 2016 conducted by Robert Faris, Hal Roberts, Bruce Etling, Nikki Bourassa, Ethan Zuckerman, and Yochai Benkler of the Berkman Klein Center for Internet and society at Harvard University investigated the bias or slant of a variety of online news sources for their coverage of the 2016 presidential election. Wikipedia specifically was identified as center-right during the 2016 presidential election, which conflicts with some of the previous studies mentioned(Faris and Benkler, 2017). The fact that this study conflicts with previous related works highlights the importance of our investigation.

The answer as to the true bias within Wikipedia's collectively-written articles is certainly still undecided and depends largely on the time period of the writing and the number of revisions the article has received. In our report, we hope to uncover more clarity as to the effect of bias on the Wikipedia articles of historically notable figures.

3 Datasets and Data Collection

The two main datasets we used for the Naive Bayes analysis were:

- https://www.ranker.com/list/the-all-time-greatest-people-in-history/alan-smithee
- https://www.ranker.com/crowdranked-list/the-all-time-worst-people-in-history

These lists are constructed through voting by users of the website and serve as a good measure of well-known famous and infamous historical figures. As a clarification, the list of the greatest people in history notes that everyone on the list is not a saint and some may have done controversial things, but the voters looked at overall positive influence. For both of these lists, we used Selenium to scrape the links to each figure's Wikipedia page. Because this page has an infinite scroll, we had to use a Selenium scroller in combination with a time buffer that progressively moved the crawler further and further down the page.

Using these Wikipedia links, we scraped each page and loaded the page's text into files for use in our Naive Bayes algorithm. We selectively chose the most important sections of the Wikipedia articles to save in the text files. For every article we only scraped the body of the article, removing unnecessary sections like metadata, reference lists, and tables, essentially anything not necessary for evaluating the core content of the article.

We also used:

- https://www.kathkyle.com/character-traits-list/

as a pre-compiled list of positive and negatively connoted words for our second version of the Naive Bayes where we use only selected words from this list as relevant tokens for calculating probabilities within the algorithm, similar to a bag-of-words model. One issue that arose with this list is that there were more positively connoted words than negative. To ensure equal representation of positive and negative words, we only selected 384/501 negative words at random from the entire list of negative words.

With these datasets, we were able to successfully run both "regular" multinomial Naive Bayes and bag-of-words Naive Bayes with only the strongly connoted words as tokens.

4 Approach, Evaluation Methods, Experiments, and Results

Text classification is not a new technique in the domain of information retrieval, and there is no shortage of other studies that analyze the neutrality of composite open-source information available online. This being said, most of these studies have a broader or completely different goal in mind, leading us to believe that our project is original due to its unique characteristics and focus on both famous and infamous historical figures in Wikipedia articles specifically. While digesting our project's intricacies, we have made some interesting observations about the applications of such a study.

We noticed our classification goals have an element of redundancy, but also that this does not necessarily diminish the validity of our findings. We identified this redundancy when noting that our project design attempts to "resolve" the classification of a historical figure from our system with the classification of the same people from a ranked list online. Looking more closely at this, we asked "why would a Wikipedia article's depiction of a person differ from this ranked list?" and found that the answer points to cultural and ideological differences between authors of different backgrounds. Essentially, we are identifying the bias of Wikipedia articles by using the rankings of a separate group of people who voted on this person's level of "famous" or "infamous[ness]". We are comparing the ideas of two groups that likely come from a very similar culture and ideological background in a global context, and also using one collaboratively created list to judge another collaboratively constructed article, which could be problematic. We can make this claim about the localization of our results by noting that the Wikipedia articles we tested on are written exclusively in English and will therefore attract mostly Western internet users who have most likely had their ideas shaped by western culture.

For the task at hand, we chose to create a classifier using multinomial Naive Bayes. There are multiple variations of the Naive Bayes classifier, all with different use cases depending on the data and the kind of classification one wants. For example, the Bernoulli Naive Bayes model works best for testing shorter documents against a small number of fea-

tures, whereas multinomial Naive Bayes allows us to classify larger documents while considering many more possible features. Given our task, we concluded that Bernoulli Bayes would not be the best model for our data, since the absence of negatively connoted language does not imply the existence of positive language, and vice versa. Both method's also use the n-1 multinomial Naive Bayes, where the program iterates through all articles, using the n-1 articles as training data and the single article left each iteration as testing data.

First, we used the simple multinomial Naive Bayes without removing stopwords or accounting for any strongly connoted words. We tested 131 famous and 132 infamous articles scraped from Ranker using Selenium and from Wikipedia using Beautiful Soup. Using the embedded links within Ranker we then were able to access the Wikipedia articles and then compiled the text content from each article into 263 separate text documents for the n-1 testing labelled as famous or infamous depending on their association to either Ranker list. Using the P(x|c) and P(c) formulas for multinomial Naive Bayes we were able to calculate the probability of a single document being either famous or infamous by considering all tokens within the article. Using P(c) and P(x|c), we compounded and multiplied these values together and outputted the class producing the highest probability. If this process produced the class name that matches the class given by Ranker.com, this test was successful in classifying the historical figure as "famous" or "infamous". Repeating this for all n documents - training on n-1 and testing for the remaining document, led to results of 242/263 correct classifications. This 92.02% accuracy shows we correctly classified most articles as either "famous" or "infamous". This outcome also suggests that Wikipedia articles about historical figures are neutral about 92.0% of the time, however, this also raises the question about the redundancy between our testing and training data as mentioned previously. We believe that our high accuracy could possibly be attributed to this redundancy because we are testing the neutrality of a description against determinations made by a group of people with a similar background and locality as the authors of Wikipedia articles.

To explore our research question further, we decided to run our multinomial Naive Bayes classifier again, but this time modifying the text documents in an intentional way in hopes this would reveal influential features of our system and experiment. After first completing the first multinomial Naive Bayes implementation and assessing its accuracy, we took a list of heavily connoted words(384 positive and 384 negative) and used these exclusively within the Naive Bayes algorithm. Further information about where this list was derived from can be found under our description of datasets used. Similar to the first method we used the scraped data from Ranker and Wikipedia, but we also applied the Porter Stemmer because we were testing exclusively for the presence of the words from our list of heavily connoted words. Without this it would have been very rare to find words of the correct case within the document that matched words from the list. We also wanted to explore, generally, how more pre-processing of our documents could affect the accuracy of our final results, either positively or negatively. We do understand that separating out the cause of a change in accuracy could not be entirely clear when we are altering multiple aspects of our testing method at the same time. We then ran the classifier again, however when calculating P(x|c) we only accounted for words that also existed within the heavily connoted word list. We hypothesized that given a high accuracy after removing words not in this list, a reasonable conclusion would describe a model in which highly connoted words are the biggest and most accurate predictor of our outcome. However, we found this decreased the accuracy of our model.

From here, we investigated whether positive and negative words had an equivalent effect on a documents' classification. For example, if a given article about Jesus Christ (used as an obviously 'famous' example) contains one occurrence of a strongly negatively connoted word, we would want to know if our model would incorrectly classify the figure as "infamous" despite many more positive word occurrences than negative in the document. To explore this, we had to parse documents by hand and investigate if many incorrectly classified documents all had particular negative or positive words in common. If many incorrectly classified documents shared a common negative or positive word, we removed this word from our curated list of 384 positive and 384

negative words to allow for a more informative classification by our model. After a few iterations of this process, we had a more evenly-weighted list of words to work with and hoped this would improve the accuracy of our model when using a bag-of-words to modify our dataset. Despite this, the accuracy of the bag-of-words model was still lower and overall had decreased from 242/263 in the traditional multinomial Naive Bayes to 132/263 with this method, slightly above that of an algorithm that randomly guesses as to an articles classification. This is an accuracy of only 50.19%. With the accuracy of this bag-of-words method being almost half as much as with our lesser-processed corpus, we believe that if we had stemmed and removed stopwords in both methods, our accuracies would show less discrepancy, but that our 92.02% accuracy may have gone down. Ultimately, we attribute this accuracy discrepancy to the lack of distinct polarity between specific positive and negative words from our list which we predict to be true for any combination of such words. The model has a negative bias toward historical figures and is more likely to label someone as "infamous" than "famous" when there exists a combination of positive and negative words in the document.

	Naive Bayes	Bucket of Words
Accurate Number of Predictions	242	132
Total Number of predictions	263	263
Accuracy	92.02%	52.09%

In another pass it would be interesting to compare results of the two methods described above on a larger dataset. We only tested on the articles taken from Ranker.com, but through the expansion of both training and testing data we might be able to get a bigger sense of what went wrong with our second method, and also help answer better the original question about sources of bias within Wikipedia. Given an unlimited amount of time for compilation, this project could be used on a much larger scale to determine discrepancies between cultural ideas and accepted norms around the globe. Using a smaller scale, as we did, allows for a more in-depth analysis of the system's behavior, but reduces the scope of our findings and possibly restricts the depth of our conclusions. Because of this, we feel comfortable with our conclusions given the scale of our data but recognize that our experiment was not comprehensive enough to decisively answer the original question.

5 Conclusions

For our first evaluation method using multinomial Naive Bayes, our accuracy of 242/263 or 92.02% was very high, enough to conclude that the language used by the collaborators accurately depicted the historical status of the people from the Ranker list. Given the previous studies and works related to this topic of Wikipedia bias, it is not surprising that these articles written about some of the most famous people in history show very low evidence of significant bias given that they have likely received some of the most significant edits when compared to the majority of Wikipedia articles. As seen in the Greenstein and Zhu paper, "Is Wikipedia Biased?", the original date of publication and number of edits received greatly impact an article's bias, so it is expected that these older articles with many edits are largely free from bias.

Our second evaluation method using Naive Bayes with only the strongly connotated words had much poorer results. We had an accuracy of 132/263 or 50.19% was only slightly greater than if we had randomly assigned infamous or famous to historical figures at random. This discrepancy in accuracy between the multinomial and connotated word Naive Bayes methods was likely due to the issue of negatively connotated words being applicable to both famous and infamous historical figures somewhat equally. With this method, we also found a tendency to over-classify historical figures as infamous, again likely caused by a higher frequency of the negative words among all documents. This could be affected and possibly remedied by using a different list of connotated words in future attempts. Alongside this, our simple method of using only single connotated words fails to account for the intent and context of the word used. However, there is another takeaway from the extremely low accuracy of this method. It is possible that this signifies a lack of biased or inflam-

matory language within these Wikipedia articles, per the initial goals of our work.

In the future, a key area of improvement would be in the bucket of words model. As touched on previously, there were several possible issues with this method, but likely most important was the failure to account for context within our connotated words. Most words intended to be seen primarily in infamous people's articles were instead found in both, and also vice versa. Obviously accounting for context is a field of NLP still at its early stages, although with new advancements like the easily accessible GPT-3 API this is increasingly a possibility for future areas of study. Also within the bucket of words model, we could potentially select connotated words using our test data as a resource and bootstrap this list as we progressively test the multitude of articles.

Another area of improvement or future study we would like to consider for both evaluation methods we used is in the selection of historical figures. It would be an interesting experiment to increase the number of Wikipedia articles we classified, specifically through the use of a crawler that branches out from one root Wikipedia article to related articles of similar figures. Using metrics on the date of publication along with the number of edits it would potentially be possible to find sources of bias and verify comparable studies like that of Greenstein and Zhu.

In future trials we would like to alter this project to work toward producing a global consensus instead of one based on western ideas and predispositions. As mentioned previously our method is inherently biased as we are using a ground truth taken from English-speakers and using Wikipedia articles written by English speakers. With a more global approach, one could potentially attempt to create an entirely new open-source encyclopedia that takes these differences into account instead of attempting to only produce "neutral" articles where the definition of "neutral" is based on and enforced by western standards. It would be interesting to explore how different cultures can lead to differences in bias due to the difference in experience of people of that culture when compared to that of English-speakers. Furthermore, the creation of a global database of ideas could lead to new domains of information retrieval research.

6 Contributions

All group members collaborated over a zoom meeting to discuss and select our topic, along with writing up the first project checkpoint

All group members met over zoom and worked on individual sections of Checkpoint 2, collaborating in real-time. Jordan and Bret wrote the project description, Amelia wrote the Related Work, Saika worked on the Data, and Frank wrote the Method description and evaluation Methodology. Although work was delegated by sections, group members were extremely collaborative and work was done in an online collaborative platform.

Saika delegated work and completed the formatting for the poster, with all other members helping to finalize the poster over a joint zoom meeting. This culminated with Jordan presenting the Introduction, Amelia presenting the Goals, Frank, and Bret presenting the Methods, and Saika presenting the Results during the poster presentation.

Jordan and Amelia delegated work and composed the initial draft of the Final Report. All other group members contributed equally to the subsequent revisions and corrections of this initial draft over google docs and overleaf.

All group members contributed equally in the planning and big picture decisions related to the writing of the software and the use of the datasets. Bret wrote the Wikipedia crawler and Ranker scraper that initially composed the bulk of our datasets. Frank then implemented the version of the Naive Bayes algorithm we used and compiled the list of connoted words for the bag-of-words dataset. All group members helped with troubleshooting and minor modifications of our state method from Project Checkpoint 2.

References

Linda A Ashtear and Shuichi Tezuka. 2020. The left-wing bias of wikipedia, Oct.

Sanmay Das and Allen Lavoie. 2014. Automated inference of point of view from user interactions in collective intelligence venues. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 82–90, Bejing, China, 22–24 Jun. PMLR.

Hal Roberts Bruce Etling Nikki Bourassa Ethan Zuckerman Faris, Robert M. and Yochai Benkler. 2017. Partisanship, propaganda, and disinformation: Online media and the 2016 u.s. presidential election. Berkman Klein Center for Internet Society Research Paper.

Shane Greenstein and Feng Zhu. 2012. Is wikipedia biased? In *Proceedings of the 26th International Conference on World Wide Web Companion*, American Economic Review, pages 102, 343–348. American Economic Association.

Shane Greenstein and Feng Zhu. 2016. Do experts or collective intelligence write with more bias? evidence from encyclopædia britannica and wikipedia. Harvard Business School.

Christoph Hube. 2017. Bias in wikipedia. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 717–721, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.